

What do different evaluation metrics tell us about saliency models?

Zoya Bylinskii*, Tilke Judd*, Aude Oliva, Antonio Torralba, and Frédo Durand

Abstract—How best to evaluate a saliency model's ability to predict where humans look in images is an open research question. The choice of evaluation metric depends on how saliency is defined and how the ground truth is represented. Metrics differ in how they rank saliency models, and this results from how false positives and false negatives are treated, whether viewing biases are accounted for, whether spatial deviations are factored in, and how the saliency maps are pre-processed. In this paper, we provide an analysis of 8 different evaluation metrics and their properties. With the help of systematic experiments and visualizations of metric computations, we add interpretability to saliency scores and more transparency to the evaluation of saliency models. Building off the differences in metric properties and behaviors, we make recommendations for metric selections under specific assumptions and for specific applications.

Index Terms—Saliency models, evaluation metrics, fixation maps, saliency applications

1 INTRODUCTION

Automatically predicting regions of high saliency in an image is useful for applications including content-aware image re-targeting, image compression and progressive transmission, object and motion detection, image retrieval and matching. Where human observers look in images is often used as a ground truth estimate of image saliency, and computational models producing a saliency value at each pixel of an image are referred to as saliency models¹.

For use in applications, dozens of computational saliency models are available to choose from [8], [9], [14], [36], but the challenge is to determine which model(s) provides the “best” approximation to human eye fixations for a given application. For example, for the input image in Fig. 1a, we include the output of 8 different saliency models in Fig. 1b. The outputs look very distinct from one another, differing in the range, location, and distribution of saliency values. As depicted in Fig. 1c, when compared to human ground truth the saliency models receive different scores according to different evaluation metrics, leaving model performance up to interpretability. How can we reconcile the difference in scores? Do these differences provide us with additional information about saliency models? What does a high score according to a particular metric tell us about a model's strengths or weaknesses?

• Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139. E-mail: {zoya, tjudd, oliva, torralba, fredo}@csail.mit.edu.
* indicates equal contribution.

1. Although the term saliency was traditionally used to refer to bottom-up conspicuity, this is no longer the case. Scene layout, object locations, and other contextual information is built into many modern saliency models.

Choosing appropriate evaluation metrics remains an open research question because this choice depends on how saliency and fixation data are defined and represented. Some metrics consider saliency at discrete fixation locations, others treat both saliency maps and ground truth fixations as distributions; some take a probabilistic approach to distribution comparison, yet others treat distributions as histograms or random variables. The inherent ambiguity in how saliency and ground truth are represented leads to different choices of metrics for reporting performance [9], [14], [53], [64], [86].

In this paper, we review other evaluation efforts (Sec. 2) and discuss issues regarding eye movement data collection and representation (Sec. 3). The representation of eye movement data as fixation locations or continuous fixation distributions impacts the subsequent selection of metrics for saliency evaluation. We analyze and visualize 8 popular evaluation metrics, discussing which attributes of saliency models each metric rewards and penalizes (Sec. 4). Our analyses and corresponding visualizations of intermediate metric computations (as in Fig. 2) add interpretability to saliency scores and more transparency to the evaluation of saliency models. Through a series of experiments systematically varying different parameters of saliency maps and the underlying ground truth, we characterize the behavior of metrics under different conditions: false positives and false negatives, center bias, and spatial deviations (Sec. 5). Finally, building off the differences in metric properties and behaviors, we make recommendations for metric selections under specific assumptions or computational constraints, and for specific applications (Sec. 6).

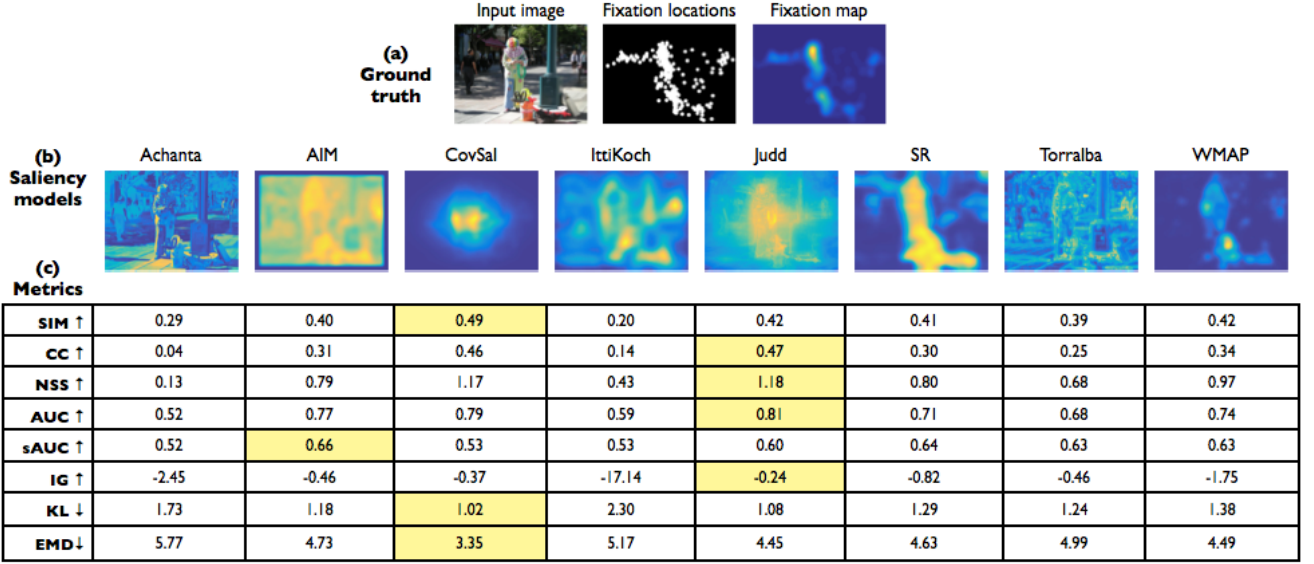


Fig. 1: Different evaluation metrics score saliency models differently. Pictured here are (b) saliency maps corresponding to 8 saliency models for the same input image. Saliency maps are evaluated on how well they approximate (a) human ground truth eye movements, represented either as discrete fixation locations or a continuous fixation map (distribution). We include the scores for each model (c) under 8 different evaluation metrics (6 similarity metrics, and 2 dissimilarity metrics), highlighting the best scoring model under each metric. This paper is about what different metrics tell us about saliency models and their fit to the human ground truth.

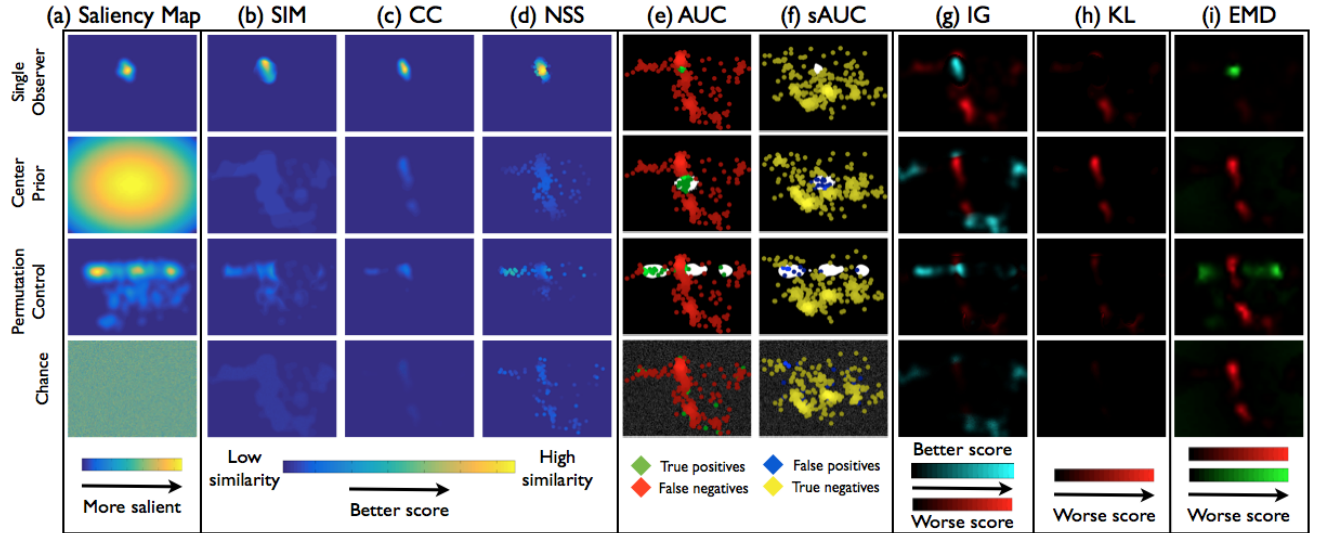


Fig. 2: A series of experiments and corresponding visualizations can help us understand what behaviors of saliency models different evaluation metrics capture. Given a natural image and ground truth human fixations on the image as in Fig. 1a, we evaluate saliency models, including the 4 baselines in column (a), at their ability to approximate ground truth. Visualizations of 8 common metrics (b-i) help elucidate the computations performed when scoring saliency models.

Saliency model	Similarity metrics						Dissimilarity metrics	
	SIM ↑	CC ↑	NSS ↑	AUC ↑	sAUC ↑	IG ↑	KL ↓	EMD ↓
Single Observer	0.38	0.53	1.65	0.80	0.64	-8.85	6.19	3.48
Center Prior	0.45	0.38	0.92	0.78	0.51	-0.43	1.24	3.72
Permutation Control	0.34	0.20	0.49	0.68	0.50	-7.33	6.12	4.59
Chance	0.33	0.00	0.00	0.50	0.50	-1.67	2.09	6.35

TABLE 1: Performance of saliency baselines (as pictured in Fig. 2) with scores averaged over MIT300 benchmark images.

2 RELATED WORK

2.1 Evaluation metrics for computer vision

Similarity metrics operating on image features have been a subject of investigation and application to different computer vision domains [47], [71], [80], [92]. Images are often represented as histograms or distributions of features, including low-level features like edges (texture), shape and color, and higher-level features like objects, object parts, and bags of low-level features. Similarity metrics applied to these feature representations have been used for classification, image retrieval, and image matching tasks [66], [70], [71]. Properties of these metrics across different computer vision tasks also apply to the task of saliency modeling, and we provide a discussion of the applications in Sec. 5. The discussion and analysis of the metrics in this paper can correspondingly be generalized to other computer vision applications.

2.2 Evaluation metrics for saliency

A number of papers in recent years have compared models across different metrics and datasets. Wilming et al. [86] discussed the choice of metrics for saliency model evaluation, deriving a set of qualitative and high-level desirable properties for metrics: “few parameters”, “intuitive scale”, “low data demand”, and “robustness”. Metrics were discussed from a theoretical standpoint without empirical experiments or quantification of metric behavior.

Le Meur and Baccino [53] reviewed many methods of comparing scanpaths and saliency maps. For evaluation, however, only 2 metrics were used to compare 4 saliency models. Sharma and Alsam [79] reported the performance of 11 models with 3 versions of the AUC metric on MIT1003 [37]. Toet [76] compared 13 saliency models on a new multiscale contrast conspicuity metric. Zhao and Koch [91] performed an analysis of saliency on 4 datasets using 3 metrics. Riche et al. [64] provided an evaluation 12 saliency models with 12 similarity metrics on Jian Li’s dataset [44]. They computed Kendall’s concordance measure between pairs of metrics to compare how metrics rank saliency models. They reported which metrics cluster together but did not discuss why.

Borji, Sihite et al. [8] compared 35 models on a number of image and video datasets using 3 metrics. Borji, Tavakoli et al. [9] compared 32 saliency models with 3 metrics for fixation prediction and additional metrics for scanpath prediction on 4 datasets. The effects of center bias and map smoothing on model evaluation were discussed. A synthetic experiment was run with a single set of random fixations while blur sigma, center bias, and border size were varied to determine how the 3 different metrics are affected by these transformations. Our analysis extends to 8 metrics tested on different variants of synthetic data to explore the space of metric behaviors.

Li et al. [45] used perceptual experiments on humans to discover a metric that most closely corresponds to the visual comparison of spatial distributions. Participants were asked to select out of pairs of saliency maps the map they perceived to be closest to the ground truth map. The human annotations were then used to order saliency models, and this ranking was compared to the rankings produced by different metrics. The authors compared 9 evaluation metrics, and used a neural network model to learn a novel metric directly from the perceptual experiment data. It is important to note, however, that distribution comparison by visual means is biased. In particular, visual comparison may be affected by the range and scale of saliency values (see Sec. 2.3).

Emami and Hoberock [21] compared 9 evaluation metrics (3 novel, 6 previously-published) in terms of human consistency. They defined the best evaluation metric as the one which best discriminates between a human saliency map and a random saliency map, as compared to the ground truth map. Human fixations were split into 2 sets, to generate human saliency maps and ground truth maps for each image. This procedure was the only criterion by which metrics were evaluated, and the chosen evaluation metric was used to compare 10 saliency models.

In this paper, we analyzed metrics commonly used in other evaluation efforts (Table 2) and appearing on the MIT Saliency Benchmark [14]. We supplemented these metrics with Information Gain (IG), recently introduced by Kümmerer et al. to address some of the issues with previous metrics [41], [42]. To visualize metric computations and highlight differences in metric behaviors, we used standard saliency models for which code is available online. These models, depicted in Fig. 1b, include Achanta [2], AIM [11], CovSal [23], IttiKoch [40], [82], Judd [37], SR [69], Torralba [77], and WMAP [49]. Models were used for visualization purposes only, as the primary focus of this paper is comparing the metrics, not the models.

This paper extends beyond tables of performance values and a literature review of metrics, to offer intuition about how metrics perform under various conditions and where they differ, using experiments with synthetic and natural data, and visualizations of metric computations. We examined the theoretical and empirical limits of metric scores, and the effects of false positives, false negatives, monotonic transformations, spatial biases, and spatial deviations on performance. This paper offers a more complete understanding of evaluation metrics and what they measure.

2.3 Qualitative evaluation of saliency

Most saliency papers include side-by-side comparisons of different saliency maps computed for the same images (as in Fig. 1b). Visualizations of saliency

Metric	Denoted here	Evaluation papers appearing in
Area under ROC Curve	AUC	[9], [21], [22], [45], [53], [64], [86], [91]
Shuffled AUC	sAUC	[8], [9], [45], [64]
Normalized Scanpath Saliency	NSS	[8], [9], [21], [45], [53], [64], [86], [91]
Pearson's Correlation Coefficient	CC	[8], [9], [21], [22], [45], [64], [86]
Earth Mover's Distance	EMD	[45], [64], [91]
Similarity or histogram intersection	SIM	[45], [64]
Kullback-Leibler divergence	KL	[21], [45], [64], [86]
Information Gain	IG	[41], [42]

TABLE 2: The most common metrics for saliency model evaluation are analyzed in this paper. We include a list of the evaluation papers (surveys of models and metrics) that report these metrics.

maps are often used to highlight improvements over previous models. A few anecdotal images might be used to showcase model strengths and weaknesses.

Bruce et al. [12] discussed the problems with visualizing saliency maps, in particular the strong effect that contrast has on the perception of saliency models. They advocated for visualizing histogram-equalized saliency maps, to make perceptual comparison more meaningful [36]. Supplementing saliency map visualizations with visualizations of metric computations (as in Fig. 2 and throughout the rest of this paper) may provide an additional means of comparison that is more tightly linked to the underlying model performance than the saliency maps themselves.

In this paper, we discuss saliency model behaviors that different metrics capture, to provide saliency researchers with quantitative support for claims about false positives and negatives, spatial biases, deviations, and other characteristics of saliency models.

3 EVALUATION SETUP

The choice of evaluation metrics should be considered in the context of the whole evaluation setup, which requires the following decisions to be made: (1) on which input images saliency models will be evaluated, (2) how the ground truth eye movements will be collected (e.g. at which distance and for how long human observers view each image), (3) how the eye movements will be represented (e.g. as discrete points, sequences, or distributions), and finally (4) which metrics will be used for comparing saliency models to the human ground truth. How the ground truth is collected and represented therefore has an impact on the metrics used at evaluation time. In this section we discuss these considerations and describe the data used in our evaluation experiments.

3.1 Data collection

Images and task:

We used the MIT300 dataset [14], [36] composed of 300 images from Flickr Creative Commons and personal collections. Eye movements were collected by allowing participants to free-view each image for 2 seconds (more details in the appendix)². Under free-viewing, longer durations do not necessarily generate new fixation patterns; observers often cycle back to previously examined image elements [22], [56], [88]. Additionally, consistency between fixations of different subjects is high just after stimulus onset but progressively decreases over time [75], [87]. Fixation data collected with a 2 second exposure is thus a reasonable testing ground for saliency models [51]. Different tasks (free viewing, visual search, etc.) also differently direct eye movements and require different model assumptions [13]. The free viewing task is most commonly used for saliency modeling as it requires fewest additional assumptions.

Eye movements:

The eye tracking set-up, including participant distance to the eye tracker, calibration error, and image size affect the assumptions that can be made about the collected data and measurement errors. Given the eye tracking set-up in the MIT300 dataset (appendix), one degree of visual angle is approximately 35 pixels. One degree of visual angle is typically used as an estimate of the size of the human fovea: e.g. how much of the image a participant sees when fixating a single point on the image. The measurement error in the MIT300 set-up is also within one degree of visual angle. This information is relevant for choosing how to represent the eye fixation data.

The robustness of the eye fixation data also depends on the number of eye fixations collected, and should inform the evaluation of computational models. In the MIT300 dataset, the eye fixations of 39 observers are available per image, more than in other datasets of similar size.

3.2 Ground truth representation

Once collected, the ground truth eye fixations can be pre-processed and formatted in a number of ways for use in saliency evaluation. There is a fundamental ambiguity in the correct representation for the fixation data, and different representational choices rely on different assumptions. One format is to encode the original fixation locations into a map with a unity value at each pixel coordinate fixated. Alternatively, the discrete fixations can be converted into a continuous distribution, a **fixation map** (Fig. 1a), by placing a Gaussian with sigma equal to one degree of visual

2. See <http://saliency.mit.edu/datasets.html> for a list of eyetracking datasets with different experimental setups, tasks, images, and exposure durations.

angle at each fixation location (a common implementation choice [53]). When formalizing the metrics in the following section, we explicitly denote the binary map of fixation locations as Q^B and the continuous fixation map (distribution) as Q^D .

Representing the fixation data as a smooth map rather than discrete fixation locations allows uncertainty in the measurements to be incorporated: error in the eye-tracking as well as the inherent uncertainty of what exactly a human observer sees when looking at a particular location on the screen. Additionally, in the case of few observers, representing the fixation data as a distribution extrapolates the data to approximate more observers, which may provide a more robust ground truth for the evaluation of saliency models. Another important consideration is that even in the case of many observers, any splitting of observer fixations in two sets will never lead to perfect overlap (due to the discrete nature of the data). In this case human consistency can not reach the upper bound of metric performance (Sec. 5.1).

On the other hand, conversion of the fixation locations into a distribution is a modification (post-processing) of the collected data, and requires selecting distribution parameters. In particular, the choice of sigma for Gaussian smoothing may significantly affect metric scores during evaluation (Sec. 4.2.1).

The assumption underlying the choice of representation is whether the underlying ground truth is a distribution from which discrete fixation locations have been sampled, or whether the fixation map is an extrapolation of discrete fixation data to the case of infinite observers. Depending on which assumptions and trade-offs can be made in a particular case (for a particular application area or dataset), these choices impact the selection of a metric.

A number of metrics for the evaluation of sequences of fixations are also available [53]. However, most saliency models and evaluations are tuned for location prediction, as sequences tend to be noisier and harder to evaluate. In the following sections, we only consider evaluations that operate on the spatial, not temporal, ground truth fixation data.

4 METRIC COMPUTATION

Different metrics have been used in different saliency evaluation efforts. In this section we describe 8 evaluation metrics with their most commonly-used implementations. Additional implementation details and variants are provided in the appendix. We include the particular advantages and disadvantages of each metric, as well as visualizations of the metric computations. Metrics can be classified in different ways, but following the discussion in Sec. 3.2 and similarly to Riche et al. [64], we categorize them as **location-based** or **distribution-based** depending on whether the ground truth is represented as discrete fixation

	Location-based	Distribution-based
Similarity metrics	AUC, sAUC, NSS, IG	SIM, CC
Dissimilarity metrics		EMD, KL

TABLE 3: Different metrics use different formats of ground truth for evaluating saliency models. Location-based metrics consider saliency map values at discrete fixation locations, while distribution-based metrics treat both ground truth fixation maps and saliency maps as continuous distributions. Good saliency models should have high values for similarity metrics and low values for dissimilarity metrics.

locations or a continuous fixation map. An additional axis of organization is whether the metrics measure a **similarity** or a **dissimilarity** between prediction and ground truth, which is relevant for interpreting metric scores. This organization is summarized in Table 3.

4.1 Location-based metrics

4.1.1 Area under ROC Curve (AUC):

Evaluating saliency as a classifier of fixations

Given the goal of predicting the fixation locations on an image, a saliency map can be interpreted as a classifier of which pixels are fixated or not. This suggests a detection metric for measuring saliency map performance. In signal detection theory, the Receiver Operating Characteristic (ROC) measures the tradeoff between true and false positives at various discrimination thresholds [30], [24]. The Area under the ROC curve, referred to as AUC, is the most widely used metric for evaluating saliency maps. The saliency map is treated as a binary classifier of fixations at various threshold values (level sets), and an ROC curve is swept out by measuring the true and false positive rates under each binary classifier (level set). Different AUC implementations differ in how true and false positives are calculated.

Computing true and false positives:

An AUC variant from Judd et al. [37], called **AUC-Judd** [14], is depicted in Fig. 3. For a given threshold, the true positive rate (**TP rate**) is the ratio of true positives to the total number of fixations, where true positives are saliency map values above threshold at *fixated pixels*. This is equivalent to the ratio of fixations falling within the level set to the total fixations.

The false positive rate (**FP rate**) is the ratio of false positives to the total number of saliency map pixels at a given threshold, where false positives are saliency map values above threshold at *unfixated pixels*. This is equivalent to the number of pixels in each level set, minus the pixels already accounted for by fixations.

Another variant of AUC by Borji et al. [7], called **AUC-Borji** [14], uses a uniform random sample of image pixels as negatives and defines the saliency map values above threshold at these pixels as false

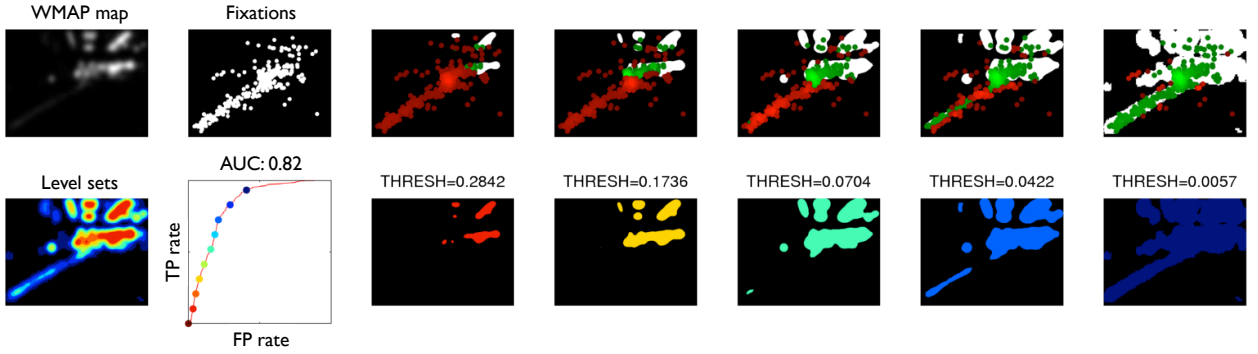


Fig. 3: The AUC metric evaluates a saliency map’s predictive power by how many ground truth human fixations it captures in successive level sets. To compute AUC, a saliency map (top left) is treated as a binary classifier of fixations at various threshold values (THRESH) and an ROC curve is swept out. Thresholding the saliency map produces the level sets in the bottom row. For each level set, the true positive rate is the proportion of fixations landing in the level set (top row, green points). The false positive rate is the proportion of image pixels covered by the level set (except for pixels on which fixations land). We include 5 level sets corresponding to 5 points on the ROC curve (with color correspondence). The final AUC score for the saliency map is the area under the ROC curve.

positives. These AUC implementations are compared in Fig. 4. The first row depicts the TP rate calculation, equivalent across implementations. The second and third rows depict the FP rate calculations in AUC-Judd and AUC-Borji, respectively. The false positive calculation in AUC-Borji is a discrete approximation of the calculation in AUC-Judd.

Sampling thresholds for the ROC curve:

The ROC curve is obtained by plotting the true positive rate against the false positive rate at various thresholds of the saliency map. Choosing how to sample thresholds to approximate the continuous ROC curve is an important implementation consideration. A saliency map is first normalized so all saliency values lie between 0 and 1. In the AUC-Judd implementation, each distinct saliency map value is used as a threshold, so this sampling strategy provides the most accurate approximation to the continuous curve. In the AUC-Borji implementation the threshold is sampled at a fixed step size (from 0 to 1 by increments of 0.1), and thus provides a suboptimal approximation for saliency maps that are not histogram equalized. For this reason, and for the otherwise similar computation to AUC-Judd, we report AUC scores using the AUC-Judd implementation in the rest of the paper. Additional implementation details and other variants of AUC are discussed in the appendix.

Compensating for center bias:

The natural distribution of fixations on an image includes a higher density near the center of an image [74]. If the first few level sets of a saliency map cover the central portion of the image, such a saliency map will often achieve a high AUC score independently of the image contents. To counter this center

bias, the shuffled AUC metric, **sAUC** [9], [74], [90], [19], [75] samples negatives from fixation locations from other images, instead of uniformly at random. In Fig. 4 the shuffled sampling strategy of sAUC (fourth row) is compared to the random sampling strategy of AUC-Borji (third row).

By averaging fixations over many images, a central Gaussian distribution naturally emerges [74], [86]. As a result, false positives sampled from other images will come predominantly from the image center. A **center prior** baseline (Fig. 2, third row), computed as a central Gaussian stretched to the image dimensions, achieves an sAUC score of 0.5. This is because at all thresholds, this baseline captures as many fixations on the current image as fixations on other images (TP rate = FP rate). Many saliency papers choose to use the sAUC metric because it compensates for the central fixation bias [74]. However, sAUC has the downside of giving more credit to off-center information (even when a central prediction is reasonable, Fig. 5) and it is recommended that it be supplemented by additional metrics [9].

Invariance to monotonic transformations:

AUC metrics measure only the relative (i.e., ordered) saliency map values at ground truth fixation locations. In other words, the AUC metrics are ambivalent to monotonic transformations. AUC is computed by varying the threshold of the saliency map and computing a trade-off between true and false positives. Lower thresholds correspond to measuring the coverage similarity between distributions, while higher thresholds correspond to measuring the similarity between the peaks of the two maps [22]. Due to how the ROC curve is computed, the AUC score for a saliency map is mostly driven by the higher thresholds: i.e., the number of ground truth fixations

Ground truth fixation locations		Saliency map level sets				
(a) TP	Fixations					
(b) FP for AUC-Judd	Non fixated locations					
(c) FP for AUC-Borji	Random samples					
(d) FP for sAUC	Shuffled samples					
Legend:		◆ True positives	◆ False negatives	◆ False positives	◆ True negatives	

Fig. 4: How true and false positives are calculated under different AUC metrics: (a) In all cases, the true positive rate is calculated as the proportion of fixations falling into the thresholded saliency map (green over green plus red). (b) In AUC-Judd, the false positive rate is the proportion of non-fixated pixels in the thresholded saliency map (blue over blue plus yellow). (c) In AUC-Borji, this calculation is approximated by sampling negatives uniformly at random and computing the proportion of negatives in the thresholded region (blue over blue plus yellow). (d) In sAUC, negatives are sampled according to the distribution of fixations in other images instead of uniformly at random. Saliency models are scored similarly under the AUC-Judd and AUC-Borji metrics, but differently under sAUC due to the sampling of false positives.

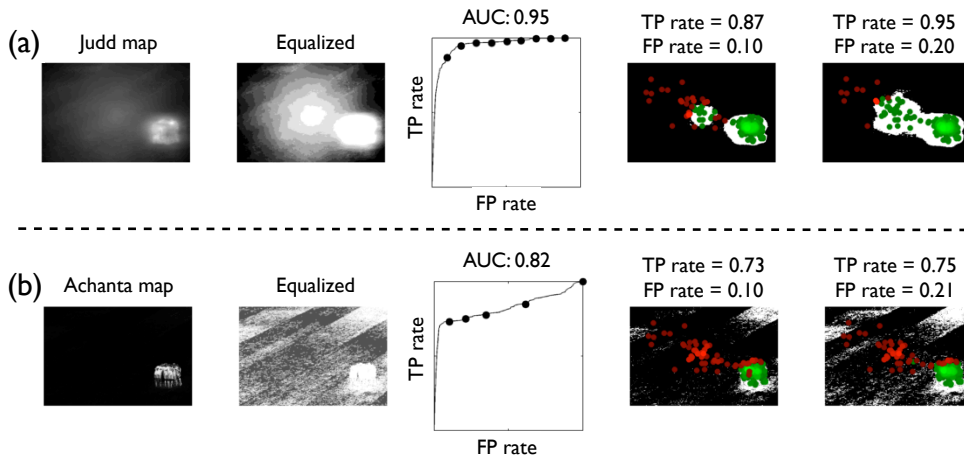


Fig. 6: The saliency map in the top row accounts for more fixations in its first few level sets than the map in the bottom row, achieving a higher AUC score overall. The AUC score is driven most by the first few level sets, while the total number of levels sets and false positives in later level sets have a significantly smaller impact. Equalizing the saliency map distributions allows us to visualize the level sets. The map in the bottom row has a smaller range of saliency values, and thus fewer level sets and sample points on the ROC curve. Both axes on the ROC curves span 0 to 1.

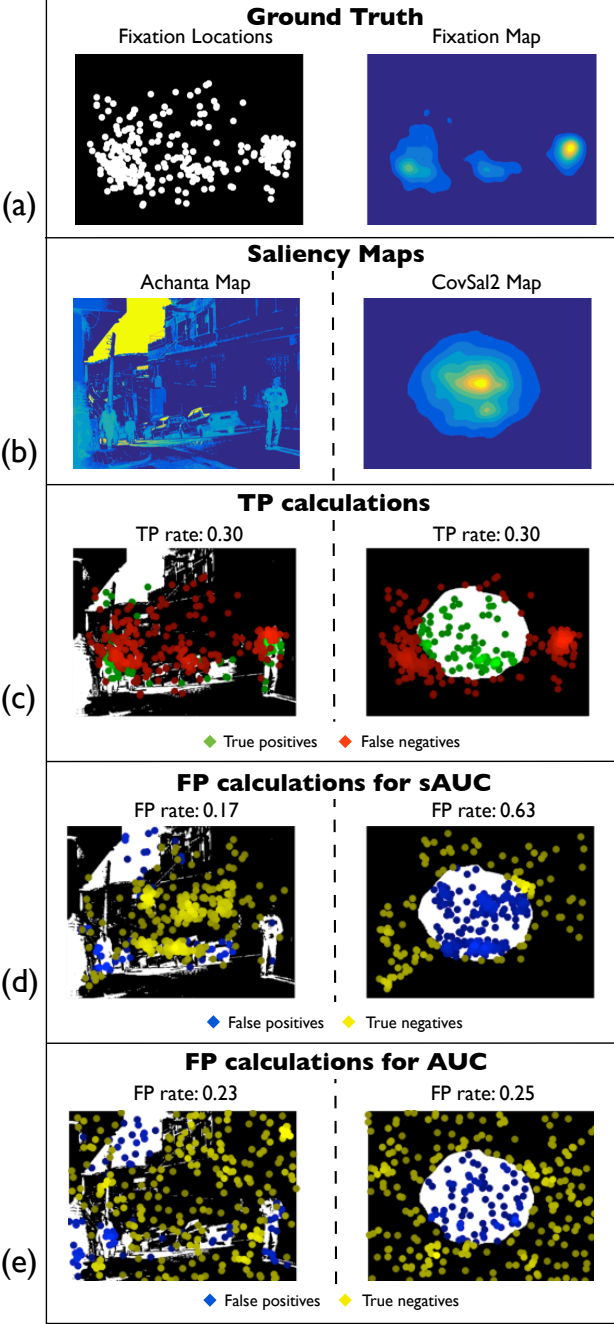


Fig. 5: An example where the sAUC metric prefers a map that makes more peripheral predictions than a center-biased map, even though the latter better predicts the ground truth. The saliency maps in (b) are compared on their ability to predict the ground truth fixations in (a). (c) For a particular level set, the true positive rate is the same for both maps. (d) The sAUC metric normalizes this value by fixations sampled from other images, more of which land in the central region of the image, thus penalizing the rightmost model for its central prediction. (e) The AUC metric, however, samples fixations uniformly at random and prefers the center-biased model which better explains the overall viewing behavior.

captured by the peaks of the saliency map (or the first few level sets as in Fig. 6). Models that place high-valued predictions at fixated locations receive high scores, while low-valued predictions at non-fixated locations are mostly ignored (Sec. 5.2). In other words, low-valued false positives are not penalized by AUC metrics, which may be unfavorable behavior for some applications (Sec. 6).

4.1.2 Normalized Scanpath Saliency (NSS): Measuring the normalized saliency at fixations

The Normalized Scanpath Saliency, **NSS** was introduced to the saliency community as a simple correspondence measure between saliency maps and ground truth, computed as the average normalized saliency at fixated locations [61]. Unlike in AUC, the absolute saliency values are part of the normalization calculation. NSS is sensitive to false positives, relative differences in saliency across the image, and monotonic transformations. Given a saliency map P and a binary map of fixation locations Q^B :

$$NSS(P, Q^B) = \frac{1}{N} \sum_i \bar{P}_i \times Q_i^B$$

$$\text{where } N = \sum_i Q_i^B \text{ and } \bar{P} = \frac{P - \mu(P)}{\sigma(P)}$$

where i indexes the i^{th} pixel, and N is the total number of fixated pixels. Chance is at 0, positive NSS indicates correspondence between maps above chance, and negative NSS indicates anti-correspondence. For instance, a unity score corresponds to fixations falling on portions of the saliency map with a saliency value one standard deviation above average.

Recall that a saliency model with high-valued predictions at fixated locations would receive a high AUC score even in the presence of many low-valued false positives (Fig. 7d). However, all false positives contribute to lowering the normalized saliency value at each fixation location, thus reducing the overall NSS score (Fig. 7c). The visualization for NSS consists of the normalized saliency value for each fixation location (i.e., \bar{P}_i where $Q_i^B = 1$).

4.1.3 Information Gain (IG):

Evaluating information gain over a baseline

Information Gain, **IG**, was recently introduced by Kümmerer et al. [41], [42], designed to handle center bias and have an interpretable linear scale. This metric measures the information gain of a saliency map over a baseline (which captures image-independent behavioral fixation biases). Given a binary map of fixations Q^B , a saliency map P , and a baseline map B , the saliency and baseline maps are rescaled and normalized to be valid probability densities and then information gain is computed as:

$$IG(P, Q^B) = \frac{1}{N} \sum_i Q_i^B [\log_2(\epsilon + P_i) - \log_2(\epsilon + B_i)]$$

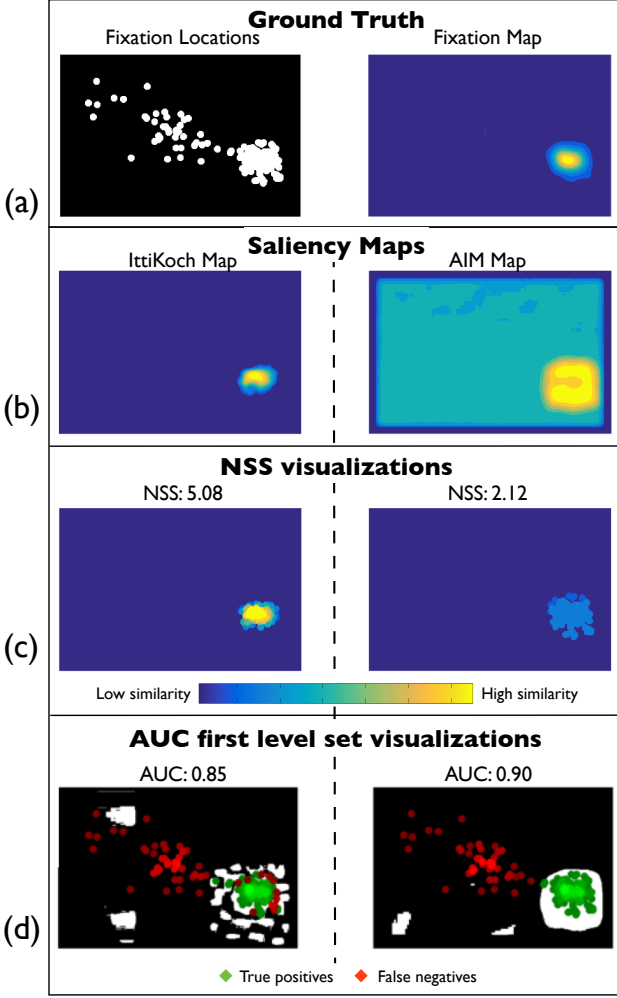


Fig. 7: An example of where AUC ignores low-valued false positives but NSS penalizes them. The saliency maps in (b) are compared on their ability to predict the ground truth fixations in (a). (c) The rightmost map is penalized by NSS for making more false positives since the normalized saliency value at fixation locations drops. (d) The AUC score of the left and right maps is very similar since a similar number of fixations fall in equally-sized level sets of the two saliency maps.

where i indexes the i^{th} pixel, N is the total number of fixated pixels, ϵ is for regularization, and information gain is measured in bits. This metric measures the average information gain of the saliency map over the baseline at fixated locations (i.e., where $Q^B = 1$). The baseline we used consists of fixations on all other images [42]. We averaged the ground truth fixation maps of all other images to create an image-independent baseline for the remaining image being evaluated. This approximates the center prior model with dataset-specific biases. A score above zero indicates the saliency map has a better prediction for the fixated locations than the baseline, beyond behavioral

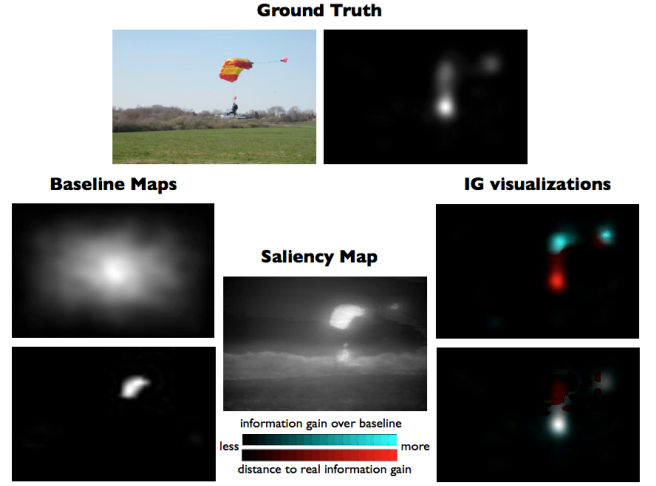


Fig. 8: We compute the information of one model over another at predicting human ground truth fixations. Here we visualize the information gain of the Judd model over an image-independent baseline (top) and the bottom-up IttiKoch1 model (bottom). Visualized in blue is the information gain over each baseline: i.e., image pixels at which the Judd model makes better predictions than each baseline. Visualized in red is the remaining distance to the real information gain: i.e., image pixels at which the Judd model underestimates saliency.

fixation biases. This formulation is amenable to replacing the baseline with any other model, and can be used for measuring the information gain of one model over another. The example in Fig. 8 contains a visualization of the information gain of the Judd model over the image-independent baseline and over the bottom-up IttiKoch model. Visualized in red are image regions for which the Judd model underestimates saliency relative to each baseline, and in blue are image regions for which the Judd model achieves a gain in performance over the baseline at predicting the ground truth. The human under the parachute can be accounted for by center bias, so the Judd model does not achieve information gain in these areas (red), but the parachute is where the Judd model outperforms the image-independent baseline (blue). On the other hand, the bottom-up IttiKoch model captures the parachute but misses the person in the center of the image, so in this case the Judd model achieves gains on the central image pixels but not on the parachute.

4.2 Distribution-based metrics

Location-based metrics achieve maximal performance precisely when all fixation locations are accounted for, with the assumption that the ground truth measurements are accurate and representative. In the extreme case of two individual observers' fixations being compared, there will be almost no overlap in the exact pixels fixated, and location-based metrics will

Attribute	Location-based	Distribution-based
Robust to measurement errors		✓
Robust to limited data		✓
Free of distribution assumptions	✓	
Ground truth not post-processed	✓	

TABLE 4: A summary of the properties of location-based and distribution-based evaluation metrics as a function of the underlying ground truth data (discrete fixation locations versus continuous fixation maps, respectively).

assign a similarity score close to zero. If this is not the desired behavior, then representing the ground truth as a distribution can make evaluation more robust. We consider another set of metrics that operate on fixation distributions instead of the original fixation locations. Following the discussion in Sec. 3.2, the advantages and disadvantages of location-based and distribution-based metrics are summarized in Table 4.

4.2.1 Similarity (SIM):

Measuring the intersection between distributions

The similarity metric, **SIM** (also referred to as *histogram intersection*), measures the similarity between two distributions, viewed as histograms. First introduced as a metric for color-based and content-based image matching [67], [73], it has gained popularity in the saliency community as a simple comparison between pairs of saliency maps. SIM is computed as the sum of the minimum values at each pixel, after normalizing the input maps. Given a saliency map P and a continuous fixation map Q^D :

$$SIM(P, Q^D) = \sum_i \min(P_i, Q_i^D) \text{ where } \sum_i P_i = \sum_i Q_i^D = 1.$$

A SIM of one indicates the distributions are the same, while a SIM of zero indicates no overlap. Fig. 9c contains a visualization of this operation. At each pixel i of the visualization, we plot $\min(P_i, Q_i^D)$. Note that the model with the sparser saliency map has a lower histogram intersection with the ground truth map. SIM is very sensitive to missing values, and penalizes predictions that fail to account for all of the ground truth density (see Sec. 5.2 for a discussion).

Effect of blur on model performance:

The downside of a distribution metric like SIM is that the choice of the Gaussian sigma (or blur) in constructing the fixation and saliency maps affects model evaluation. For instance, as demonstrated in the synthetic experiment in Fig. 12a, even if the correct location is predicted, SIM will only reach its maximal value when the saliency map’s sigma exactly matches the ground truth sigma. The SIM score drops off drastically under different sigma values, more drastically than the other metrics. Fine-tuning this optimal blur value on a training set with similar parameters as the

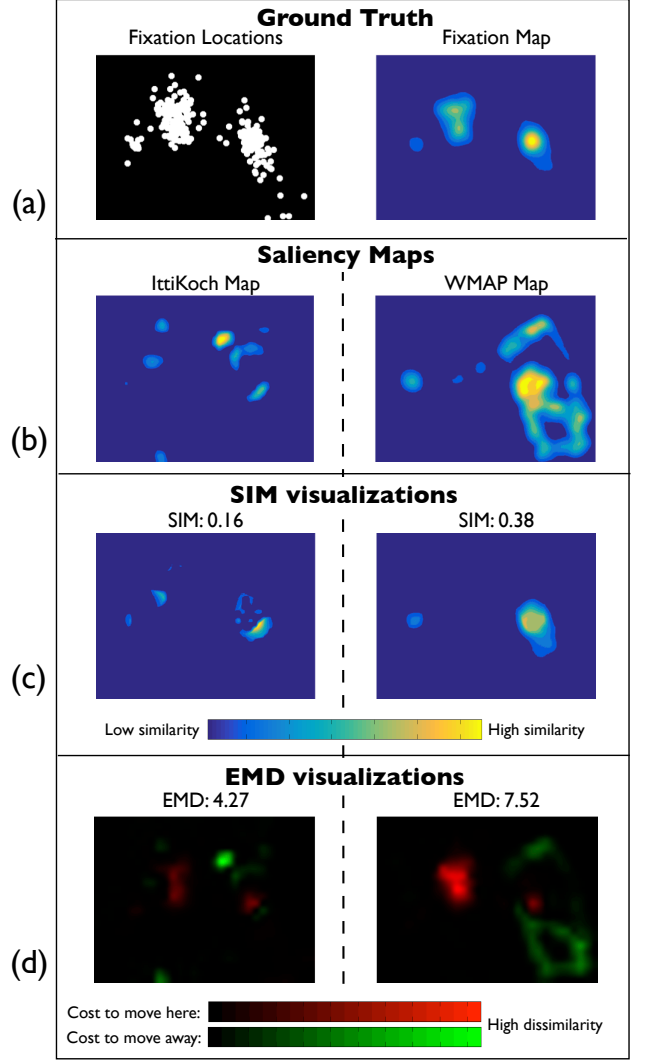


Fig. 9: An example where EMD prefers sparser predictions, even if they do not perfectly align with ground truth fixation locations, while SIM penalizes misalignment. The saliency maps in (b) are compared on their ability to predict the ground truth fixation map in (a). (c) The saliency map on the left makes sparser predictions and thus has a smaller area of intersection with the ground truth density than the model on the right. (d) The map on the left has a better EMD score because the predicted density is not spatially far from the ground truth density, while the map on the right needs a lot of density moved to match the ground truth.

test set (eyetracking set-up, viewing angle) can help inflate model performances [14], [36].

The SIM metric is good for evaluating partial matches, where a subset of the saliency map accounts for the ground truth fixation map. As a side-effect, false negatives tend to be penalized more than false positives. For other applications, however, a metric that treats false positives and false negatives symmetrically may be preferred.

4.2.2 Pearson's Correlation Coefficient (CC): Evaluating the linear relationship between distributions

The Pearson's Correlation Coefficient, **CC**, also called *linear correlation coefficient* treats the saliency and fixation maps, P and Q^D , as random variables and measures the linear relationship between them [54]. This is meaningful in the context of comparing the relative saliency values at different image regions, for instance for scanpath analysis. CC is computed as:

$$CC(P, Q^D) = \frac{\sigma(P, Q^D)}{\sigma(P) \times \sigma(Q^D)}$$

where $\sigma(P, Q^D)$ is the covariance of P and Q^D . CC is symmetric and penalizes false positives and negatives equally. It is invariant to linear (but not arbitrary monotonic) transformations. High positive pixel-wise CC values occur where both the saliency map and ground truth fixation map have values of similar magnitudes at the same locations. In Fig. 10 is an illustrative example comparing the behaviors of SIM and CC: where SIM penalizes false negatives significantly more than false positives, but CC treats both false positives and negatives symmetrically. For the visualization of CC in Fig. 10d, each pixel i has value:

$$\frac{P_i \times Q_i^D}{\sqrt{\sum_j (P_j^2 + (Q_j^D)^2)}}$$

Due to its symmetric computation, CC can not distinguish whether differences between maps are due mainly to false positives or false negatives. Other metrics may be preferable if this kind of analysis is of interest.

4.2.3 Kullback-Leibler divergence (KL): Evaluating saliency with a probabilistic interpretation

Kullback-Leibler (**KL**) takes a probabilistic interpretation of the saliency and fixation maps, evaluating the loss of information when distribution P (the saliency map) is used to approximate distribution Q^D (the ground truth fixation map). Formally:

$$KL(P, Q^D) = \sum_i Q_i^D \log \left(\epsilon + \frac{Q_i^D}{\epsilon + P_i} \right)$$

where ϵ is a regularization constant. KL is an asymmetric dissimilarity metric, with a lower score indicating a better approximation of the ground truth by the saliency map. We compute a per-pixel score to visualize the KL computation (Fig. 11d). For each pixel i in the visualization, we plot $Q_i^D \log \left(\epsilon + \frac{Q_i^D}{\epsilon + P_i} \right)$. Wherever the ground truth value Q_i^D is non-zero but P_i is close to or equal to zero, a large quantity is added to the KL score. Such regions are the brightest in the KL visualization. There are more bright regions in the rightmost map of Fig. 11d, corresponding to areas in the ground truth map that were left unaccounted for by the predicted saliency.

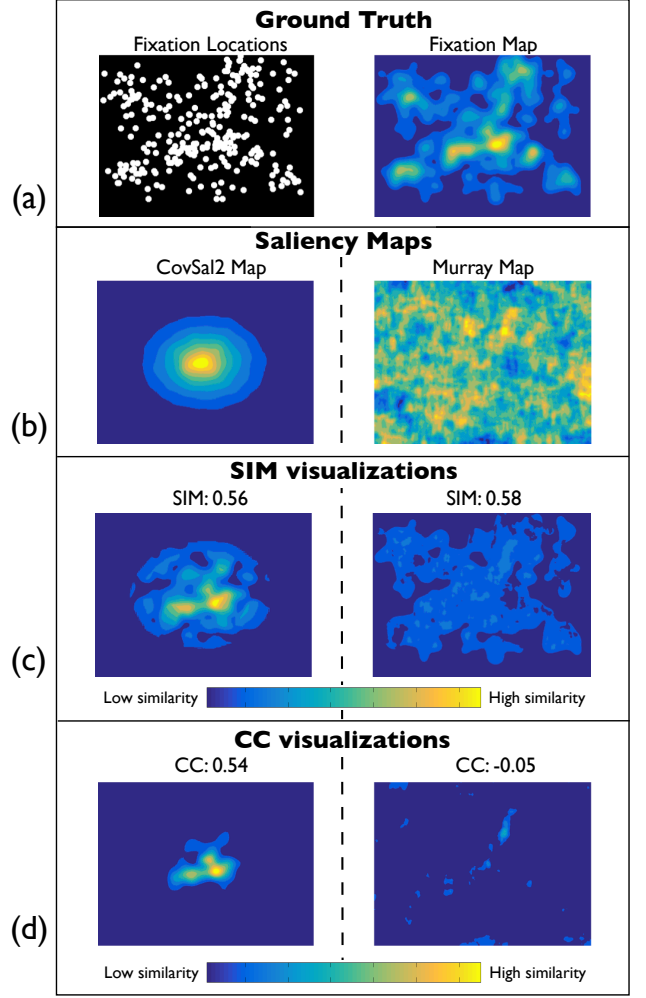


Fig. 10: An example where CC treats false positives and negatives symmetrically, but SIM places less emphasis on false positives than false negatives. The saliency maps in (b) are compared on their ability to predict the ground truth fixation map in (a). (c) Both saliency maps have similar SIM scores, (d) but the saliency map on the right has a lower CC score because false positives lower the overall correlation.

Both models compared in Fig. 11 are image-agnostic: one is a chance model that assigns a uniform value to each pixel in the image, and the other is a **permutation control** model which randomly selects a fixation map from another image. The latter model is more likely to capture viewing biases common across images. The permutation control model is above chance for many of the metrics in Table 1. However, KL is so sensitive to zero-values that a sparse set of predictions is penalized very harshly, significantly worse than chance.

4.2.4 Earth Mover's Distance (EMD): Incorporating spatial distance into evaluation

All the metrics discussed so far have no notion of how spatially far away the prediction is from the ground

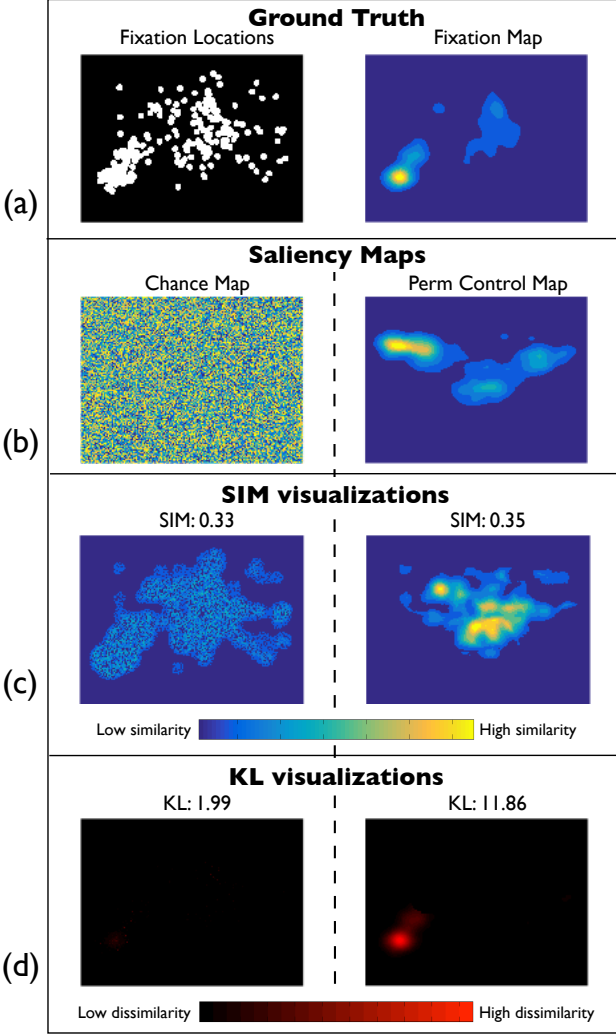


Fig. 11: An example demonstrating the sensitivity of KL to false negatives. The baseline saliency maps in (b) are compared on their ability to predict the ground truth fixation map in (a). Both saliency maps are image-agnostic, but assign saliency values differently. (c) They receive similar scores under the SIM metric. (d) However, because the chance map places uniformly-sampled saliency values at all image pixels, it contains fewer zero values, and is favored by KL (low KL divergence can not even be seen on the visualization). The rightmost map samples saliency from another image, and contains zero-values at multiple fixated locations in the current image. It is highly penalized by KL which is particularly sensitive to false negatives.

truth. Any map that has no pixel overlap with the ground truth will receive the same score, no matter how its predictions are distributed (Fig. 12b). Incorporating a measure of spatial distance can broaden comparisons, and allow for graceful degradation when the ground truth measurements have position error.

The Earth Mover’s Distance, **EMD** [67] [59], measures the spatial distance between two probability distributions over a region. Computationally, it is the minimum cost of morphing one distribution into the

other. This is visualized in Fig. 9d where in green are all the saliency map locations from which density needs to be moved, and in red are all the fixation map locations where density needs to be moved to. The total cost is the amount of density moved times the distance by which it is moved, and corresponds to brightness of the pixels in the visualization. We used the following linear time variant of EMD [59]:

$$\widehat{EMD}(P, Q^D) = (\min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij}) + |\sum_i P_i - \sum_j Q_j^D| \max_{i,j} d_{ij}$$

$$s.t. f_{ij} \geq 0 \quad \sum_j f_{ij} \leq P_i, \quad \sum_i f_{ij} \leq Q_j^D,$$

$$\sum_{i,j} f_{ij} = \min(\sum_i P_i, \sum_j Q_j^D)$$

where each f_{ij} represents the amount of density transported from the i th supply to the j th demand and d_{ij} is the *ground distance* between bin i and bin j in the distribution. A larger EMD indicates a larger difference between two distributions while an EMD of zero indicates that two distributions are the same.

Generally, saliency maps that spread density over a larger area have larger EMD values (i.e., worse scores) as all the extra density has to be moved to match the ground truth map (Fig. 9). EMD penalizes false positives proportionally to the spatial distance they are from the ground truth (Sec. 5.2).

4.3 Normalization of saliency maps

Metric computations often involve normalizing the input maps. This allows maps with different saliency value ranges to be compared. A saliency map S can be normalized in a number of ways:

(a) **Normalization by range:** $S \rightarrow \frac{S - \min(S)}{\max(S) - \min(S)}$

(b) **Normalization by variance**³: $S \rightarrow \frac{S - \mu(S)}{\sigma(S)}$

(c) **Normalization by sum:** $S \rightarrow \frac{S}{\text{sum}(S)}$

Table 5 lists the normalization strategies applied to saliency maps by the metrics in this paper. Another approach is **normalization by histogram matching**, with histogram equalization being a special case. Histogram matching is a monotonic transformation that remaps (re-bins) a saliency map’s values to a new set of values such that the number of saliency values per bin matches a target distribution. Histogram matching does not affect AUC calculations⁴, but does affect all the other metrics. Histogram matching can make a saliency map more peaked or more uniform. This has

3. This is also often called standardization.

4. Unless the thresholds for the ROC curve are not adjusted accordingly. For instance, in the ROC-Borji implementation with uniform threshold sampling, histogram matching changes the number of saliency map values in each bin (at each threshold).

Metric	Normalized by range	Normalized by variance	Normalized by sum
AUC	✓		
sAUC	✓		
NSS		✓	
CC		✓	
EMD			✓
SIM			✓
KL			✓
IG			✓

TABLE 5: Different metrics use different normalization strategies for pre-processing saliency maps prior to scoring them. Normalization can change the extent to which the range of saliency values and outliers affect performance.

different effects on metrics: for instance, EMD prefers sparser maps provided the predicted locations are near the target locations - the less density to move, the better. However, more widely-distributed predictions are more likely to have non-zero values at the target locations and thus better scores on the other metrics. These are important considerations for designing and preprocessing saliency maps (the appendix includes a discussion of how this can affect evaluation).

Different normalization schemes can also change how metric scores are impacted by very high and very low values in a saliency map. For instance, in the case of NSS, if a large outlier saliency value occurs at least at one of the fixation locations, then the resulting NSS score will be correspondingly high (since it is an outlier, it will not be significantly affected by normalization). Alternatively, if most saliency map values are large and positive except at fixation locations, then the normalized saliency map values at the fixation locations can be arbitrarily large negative numbers.

5 ANALYSIS OF METRIC BEHAVIOR

Metric behavior is a function of implementation considerations including whether the metrics operate on fixation locations or continuous distributions, how they preprocess the input maps, and how the per-pixel scores are computed, as discussed in Sec. 4 and summarized in Table 6. In this section, building upon differences in metric computation, we present a set of analyses and experiments to quantify these differences in metric behavior.

5.1 Empirical limits of metrics

One of the differences between location-based and distribution-based metrics is that the empirical limits of location-based metrics (AUC, sAUC, NSS, IG) on a given dataset do not reach the theoretical limits (Table 7). The sets of fixated and non-fixated locations are not disjoint, and thus no classifier can reach its theoretical limit [86]. In this regard, the distribution metrics are more robust. Although different sets of observers fixate similar but not identical locations,

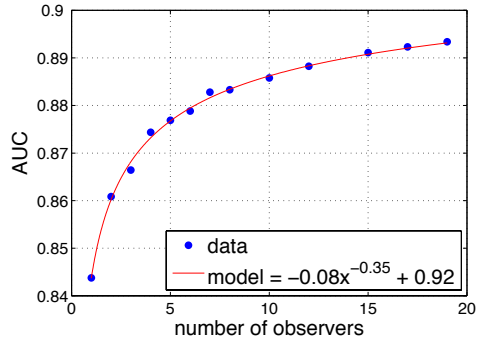


Fig. 13: We obtain an upper bound on saliency model performance by considering how well humans predict other humans in the MIT300 dataset. We plot the AUC-Judd scores when the fixations of n observers are used to predict the fixations of another n observers, for increasing n . Based on extrapolation of the power curve that fits the data, the limit of human performance is 0.92 under AUC-Judd.

continuous fixation maps converge to the same underlying distribution as the number of observers increases. To make scores comparable across metrics and datasets, empirical metric limits can be computed. Empirical limits are specific to a dataset, dependent on the consistency between humans, and can be used as a realistic upper bound for model performance. We measured human consistency using the fixations of one group of observers to predict the fixations of another group. By increasing the number of observers, we extrapolated the performance to infinite observers (other ways of defining and computing human consistency are discussed in the appendix).

For example, Fig. 13 plots the AUC-Judd score as the number of observers used for the ground truth fixation map and the number of observers used for prediction increases from $n = 1$ to $n = 19$ (half of the total 39 observers). We fit these points to the power function $f(n) = a*n^b + c$, constraining b to be negative and c to lie within the valid theoretical range of the metric. For AUC-Judd, the extrapolated performance in the limit of infinite observers ($n \rightarrow \infty$) is $c = 0.92$. The empirical limits and 95% confidence bounds for all the metrics are listed in Table 7.

Once the empirical limit has been computed for a given metric on a given dataset, this limit can be used to normalize the scores for all computational models under this metric [61].

5.2 Treatment of false positives and negatives

As we have seen in the previous sections, different metric computations place different weights on the presence of false positives and negatives in the predicted saliency relative to the ground truth. To directly compare the extent to which different metrics penalize false negatives, we performed a series of systematic tests. Starting with the ground truth fixation map, we progressively removed

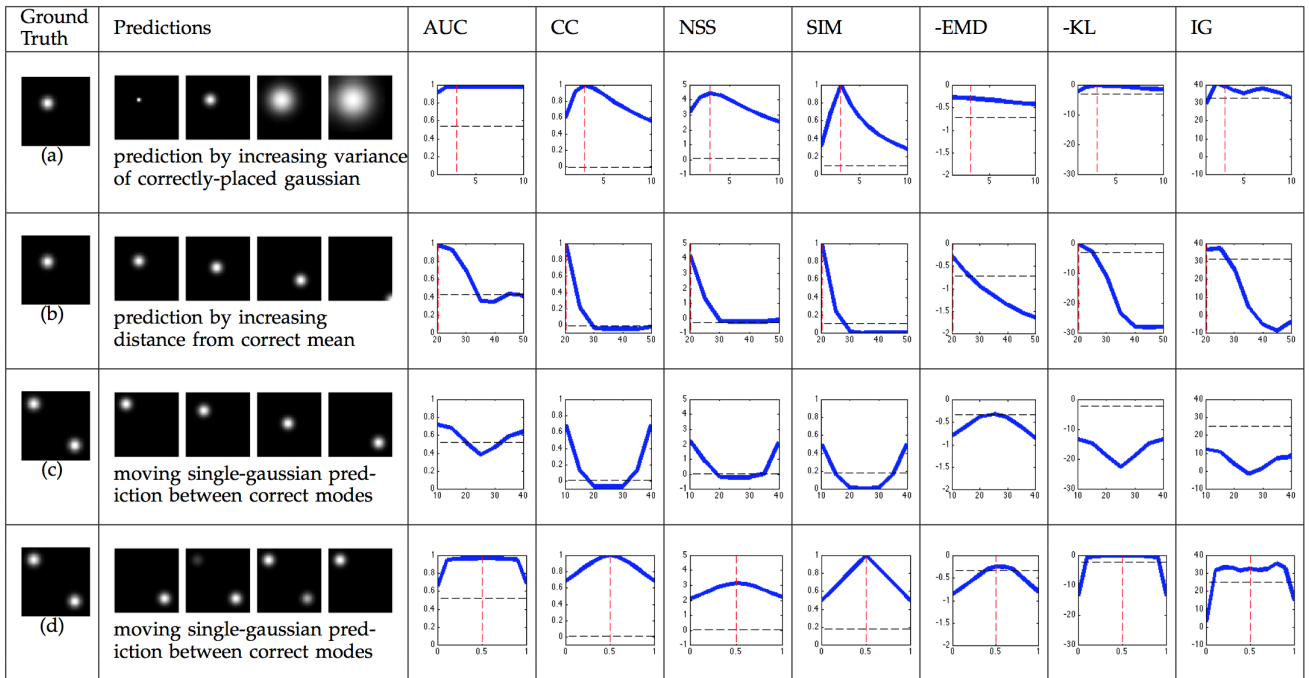


Fig. 12: We systematically varied parameters of a saliency map in order to quantify effects on metric scores. Each row corresponds to varying a single parameter value of the prediction: (a) variance, (b-c) location, and (d) relative weight. The x-axis of each subplot spans the parameter range, with the dotted red line corresponding to the ground truth parameter setting (if applicable). The y-axis is different across metrics but constant for a given metric. The dotted black line is chance performance. EMD and KL y-axes have been flipped so a higher y-value indicates better performance across all subplots.

	AUC	sAUC	SIM	CC	KL	IG	NSS	EMD
Implementation								
Bounded	✓	✓	✓	✓				
Location-based, parameter-free	✓	✓				✓	✓	
Local computations, differentiable			✓	✓	✓	✓	✓	
Symmetric			✓					✓
Metric space					✓	✓		✓
Behavior								
Invariant to monotonic transformations	✓	✓						
Invariant to linear transformations (scale)	✓	✓		✓				
Discounts viewing bias		✓				✓		
Most affected by false negatives			✓		✓	✓		
Scales with spatial distance								✓

TABLE 6: Properties of the 8 evaluation metrics considered in this paper.

different amounts of salient pixels⁵, and evaluated the similarity of the resulting map to the original ground truth map under different metrics. We measured the drop in score with 25%, 50%, and 75% false negatives. To make comparison across metrics possible, we normalized this change in score by the score difference between the upper limit and chance for each metric. For convenience, we call this the **chance-normalized score**. For instance, for the AUC-Judd metric the upper limit is 0.92, chance is at 0.50, and the score with 75% false negatives is 0.67. The chance-normalized score is: $100\% \times (0.92 - 0.67)/(0.92 - 0.50) = 60\%$. Values for the other metrics are available in Table 8.

5. Pixels with a saliency value above the mean saliency map value were selected uniformly at random and set to 0.

KL, IG, and SIM are most sensitive to false negatives: Under the definition of KL, if the prediction is close to zero where the ground truth has a non-zero value, the per-pixel KL divergence values can grow arbitrarily large. This means that KL penalizes models with false negatives significantly more than it penalizes models with false positives. In Table 8, KL scores drop below chance with only 25% false negatives. As another illustrative example, consider the single observer model where one observer’s fixations are used to predict the fixations of the remaining $n-1$ observers (Table 1). In general, a single observer’s fixations capture a subset of the locations fixated by a population of observers. This model is the worst-performing baseline model under the KL metric (KL

Metric limits	Similarity metrics						Dissimilarity metrics	
	AUC \uparrow	sAUC \uparrow	NSS \uparrow	SIM \uparrow	CC \uparrow	IG \uparrow	EMD \downarrow	KL \downarrow
Theoretical range (best score in bold)	[0,1]	[0,1]	$[-\infty, \infty]$	[0,1]	[-1,1]	$[-\infty, \infty]$	[0, ∞]	[0, ∞]
Empirical limit (with 95% confidence bounds)	0.92 (0.91; 0.93)	0.81 (0.79; 0.83)	3.29 (3.08; 3.50)	1 (0.76; 1.24)	1 (0.82; 1.18)	1.80 (1.59; 2.00)	0 0	0 0

TABLE 7: Different metric scores span different ranges, while the empirical limits of the metrics are specific to a dataset. Taking into account the theoretical and empirical limits makes model comparison possible across metrics and across datasets. An empirical limit is the performance achievable on this dataset by comparing humans to humans. It is calculated by computing the score when n observers predict another n observers, with n taken to the limit by extrapolating empirical data. Included are upper limits for the similarity metrics and lower limits for the dissimilarity metrics.

Map	EMD \downarrow	CC \uparrow	NSS \uparrow	AUC \uparrow	SIM \uparrow	IG \uparrow	KL \downarrow
Orig	0.00 (0%)	1.00 (0%)	3.29 (0%)	0.92 (0%)	1.00 (0%)	1.79 (0%)	0.00 (0%)
-25%	0.13 (2%)	0.85 (15%)	2.66 (19%)	0.85 (17%)	0.78 (33%)	-2.21 (116%)	2.55 (122%)
-50%	0.16 (3%)	0.70 (30%)	2.18 (34%)	0.77 (36%)	0.59 (61%)	-6.78 (247%)	5.64 (270%)
-75%	1.09 (17%)	0.50 (50%)	1.57 (52%)	0.67 (60%)	0.45 (82%)	-11.1 (372%)	8.18 (391%)

TABLE 8: Different metrics have different sensitivities to false negatives. The metrics are sorted in order of increasing sensitivity with 25%, 50%, and 75% false negatives, showing that EMD is least affected and KL is most affected. Scores are averaged over all 300 MIT benchmark fixation maps. Below each score is the percentage drop in performance from the metric’s limit, normalized by the percentage drop to chance level. For instance, KL penalizes false negatives so harshly that the drop in performance with 75% false negatives is almost 400% worse than chance.

= 6.19), worse than chance (KL = 2.09). Due to KL’s normalization calculation which does not discount a map with uniform saliency values, predicting a unity value at all pixels has a better KL score (1.48). Note that due to the analogous computation of the IG metric, its behavior is similar to KL on these baselines.

After KL and IG, SIM is most affected by false negatives (Table 8). All three metrics assign a lower score to the single observer model than the center prior model (Table 1). This is because central predictions are likely to capture the viewing patterns of a larger number of observers. False positives enter the SIM calculation during the normalization step, driving the saliency values down and decreasing the histogram intersection. However, each false negative has a larger impact on histogram overlap.

AUC ignores low-valued false positives:

The scores for the AUC metrics depend on which level sets the false positives fall into, where the false positives that contribute to the first few level sets are penalized most, but the false positives contributing to the last level set do not have a large impact on performance. This is why we don’t see a huge penalty incurred by the model with many low-valued false positives in Fig. 7. Under AUC, saliency maps that place different amounts of density at the correct (fixated) locations will receive a similar score (Fig. 12d).

NSS and CC are equally affected by false positives and negatives: During the normalization step of NSS, a few false positives will be washed out by the other saliency values and will not significantly affect the saliency values at fixation locations. However, as the number of false positives increases, they begin to have a larger influence on the normalization calculation, driving the overall NSS score down.

The formulation of CC makes explicit that it has a symmetric treatment of false positives and negatives. Note however that NSS is highly related to CC, and can be viewed as a discrete approximation (see the appendix). NSS behavior will be very similar to CC, including treatment of false positives and negatives.

EMD’s penalty depends on spatial distance: EMD is least sensitive to uniformly-occurring false negatives (as in the experiment in Table 8) because the EMD calculation can redistribute saliency values from nearby pixels to compensate. It does not cost much to move density to nearby pixels that may be missing density. However, false negatives that are spatially far away from any predicted density are highly penalized. Similarly, EMD’s penalty for false positives depends on their spatial location relative to the ground truth, in that false positives close to ground truth locations can be redistributed to those locations at low cost, but distant false positives are highly penalized (Fig. 9).

5.3 Systematic viewing biases

Common to many images is a higher density of fixations in the center of the image compared to the periphery, a function of both photographer bias (i.e., centering the main subject) and observer viewing biases. The effect of center bias on model evaluation has received much attention [13], [20], [43], [57], [63], [74], [75], [91]. In this section we discuss center bias in the context of the metrics in this paper.

A center prior reduces the number of false negatives: From Table 1, we see that the center model achieves a performance boost over chance according to all metrics except sAUC. It also leads to fewer false negatives when predicting the fixations of a population than a single observer model, resulting in higher SIM, KL, and IG scores (since these metrics are most sensitive to false negatives).

Discounting the center can negatively bias predictions: The sAUC metric attempts to compensate for viewing bias by sampling negatives from other images, which in the limit of many images and many samples corresponds to sampling negatives from a central Gaussian. As a result, the center baseline is at chance on this metric. sAUC has the downside of giving disproportionately more credit to off-center information [9]. In particular, for an image with a strong central viewing bias, both positives and negatives would come from the same image region, and a correct prediction would be at chance (Fig. 5).

IG can provide a direct comparison to center bias: Information gain over an image-independent baseline provides a more intuitive way to interpret model performance relative to center bias. If a model can not explain fixation patterns on an image beyond systematic viewing biases, such a model will have 0 gain over a center model.

EMD spatially hedges its bets: The EMD metric prefers models that hedge their bets if all the ground truth locations can not be accurately predicted (Fig. 12c). For instance, if an image is fixated in multiple locations, the EMD metric will favor a prediction that falls spatially between the fixated locations instead of one that captures a subset of the fixated locations (contrary to the behavior of the other metrics).

A simple center prior predicts the center of the image as most salient. It is a good approximation of average viewing behavior on images under laboratory conditions, where an image is projected for a few seconds on a computer screen in front of an observer [5]. A dataset-specific center model can be achieved by averaging fixation maps over a large set of images.

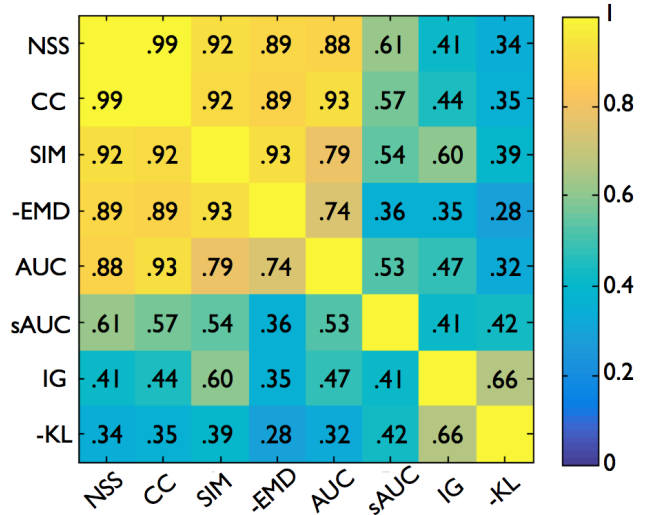


Fig. 14: Pairwise correlation between metrics at ranking saliency models on the MIT300 benchmark. The EMD and KL dissimilarity scores have been inverted to make them comparable to the similarity metrics. The first 5 metrics listed are highly correlated with each other. Of the remaining metrics, KL and IG are most highly correlated with each other and most uncorrelated with all the other metrics.

Knowing nothing about image content, it makes sense to capture average fixation behavior (the center model is a simple prior). Overall, if the goal is to predict natural viewing behavior on an image, center bias is part of the viewing behavior and discounting it entirely may be suboptimal. However, knowing why different metrics may favor center-biased models is important for interpreting model scores.

5.4 Relationship between metrics

As the main application area for the saliency metrics discussed in this paper is to rank saliency models, we can measure how correlated the model rankings are across these metrics. In Fig. 14 is the pairwise Spearman correlation matrix for our 8 metrics. The pairwise Spearman correlations between the five metrics NSS, CC, AUC, EMD, and SIM range from 0.76 to 0.98. Because these 5 metrics are highly correlated with each other, we refer to them as the **similarity cluster** of metrics. CC and NSS are most highly correlated due to their analogous computations (see appendix). Among the remaining metrics, KL and IG also have a relatively high correlation due to their related computations. SIM is more closely related to KL and IG than to the similarity cluster of metrics.

Driven by extreme sensitivity to false negatives, KL, IG, and SIM rank saliency models differently than the similarity cluster. As discussed in Sec. 5.2 and depicted in Table 1, the single-observer model is worse than chance according to these metrics, but is a top-performing model across the other metrics.

Depending on the application area, this might be sub-optimal behavior. At the same time, KL and IG have a natural probabilistic interpretation and are appropriate in cases where missing any ground truth fixation locations should be highly penalized (Sec. 6.2).

Although EMD is the only metric that takes into account the spatial distance between distributions, it nevertheless ranks saliency models similarly as the other metrics in the similarity cluster. This is likely the case for two reasons: EMD is center biased like the other metrics in the similarity cluster (see Table 1 and the discussion in Sec. 5.3) and the types of mistakes saliency models make are less often driven by a shift of the prediction (imprecise localization) than by completely incorrect predictions.

Shuffled AUC has low correlations with the other metrics because of the way it modifies how predictions at different spatial locations on the image are treated. A model with more central predictions will be ranked lower according to sAUC than a model with more peripheral predictions (Fig. 5). For these reasons, sAUC has been disfavored by some evaluations [12], [45], [53]. Of relevance is also a decision of whether the center bias issue should be handled on the part of the dataset, the model, or the metric. Center bias has been found in different types of image and video datasets, across image types and even observer tasks [8], [9], [15], [17], [33], [35]. Entirely eliminating center bias from the dataset might prove difficult. An alternative is optimizing models to include a center bias [36], [37], [41], [42], [58], [91]. In this case, the metric should be ambivalent to any model or dataset biases. The saliency metrics are much more correlated once center bias is accounted for by the models [41], [42].

5.5 Comparisons to related work:

The metric correlations reported in the last section are in agreement with Riche et al. [64] who correlated metric scores on another saliency dataset and found that KL and sAUC are most different from the other metrics. The other metrics evaluated, including AUC, CC, NSS, and SIM formed a single cluster, and they combined their rankings to produce a summary metric called *Cluster*. The replication of these results on the MIT300 dataset is a promising result showing that metric performances generalize across datasets.

Bruce et al. [12] favored AUC metrics in their analyses due to insensitivity of AUC to the absolute numeric values or contrast of saliency maps.

Qi and Koch [91] performed a saliency evaluation using the AUC metric due to its popularity and wide-use in the saliency community. To offset the bias AUC has for true positives relative to the false positives, they included NSS and EMD scores during evaluation.

Emami and Hoberock [21] used human consistency as a selection criteria to compare 9 metrics. They found that NSS and CC best discriminate between human saliency maps and random saliency maps. Under the same criterion, they listed KL as the worst metric. This is similar to comparing the metrics by their ranking of the baseline models in Table 1.

Li et al. [45] used crowd-sourced experiments to measure which metric best corresponds to human perception. They found that perception-based ranking most closely matched that of NSS, CC, and SIM, and was furthest from KL and EMD. The authors noted that human perception was driven by the most salient locations, the compactness of salient locations (i.e., low false positives), and a similar number of salient regions as the ground truth. Because qualitative comparison is often included in saliency papers, with saliency maps plotted side-by-side, a metric that matches perceptual similarity will more intuitively correspond to the plots if no additional visualizations are provided. However, as discussed in Sec. 2.2, human perception when comparing distributions might be biased in various ways.

We propose to extend the criteria when choosing metrics to account for different metric behaviors and properties (Table 6). Moreover, the choice of metrics may not be universal across applications. Based on all the analyses provided in this paper, in the next section we provide recommendations for choosing metrics for different applications.

6 SELECTING METRICS

Given the properties and behaviors of metrics discussed in the previous sections, here we consider use cases where different metrics might be most appropriate for evaluating saliency models. Metric choices depend on the underlying assumptions and constraints of an application area, under which different metric properties might be favorable.

6.1 Model comparison on fixation prediction

On standard fixation prediction datasets [14], the goal is to rank models at their ability to predict human ground truth fixations as best as possible. The Normalized Scanpath Saliency (NSS) and Pearson’s Correlation Coefficient (CC) metrics both provide a balanced treatment of false positives and false negatives. Choosing between them requires choosing between representing ground truth as fixation locations or continuous fixation distributions, respectively, with different tradeoffs (Table 4). Note that CC and NSS will measure a model’s overall ability to predict fixation behavior, including any systematic viewing biases. If the goal is to measure a model’s ability to predict fixation behavior beyond viewing biases, then the Information Gain (IG) metric may be more appropriate. It remains an open question as to whether

center bias should be accounted for at the time of model computation or model evaluation. If all models equally accounted for center bias, then model ranking according to IG and CC or NSS would be highly correlated [41], [42]. Since different models are based on different assumptions, we recommend reporting both IG and one of CC and NSS.

For historical reasons, and for facilitating comparison to existing saliency literature, an ROC analysis may also be included. Additional computational limitations may cause one metric to be favored over another - for instance, a saliency model may need to be optimized with respect to a differentiable metric, or it may be infeasible for metric calculation to be a computational bottleneck (Table 6). To demonstrate a model's superiority under specific conditions (e.g. spatial deviations), or for specific applications (Sec. 6.2) a different choice of metrics may be more appropriate.

6.2 Other saliency applications

The evaluation of saliency models for different applications may call for different metrics. Extensive lists of saliency applications are available in [6], [34]. Here, we only include a few illustrative examples.

Detection:

In detection applications using saliency, such as object and motion detection, surveillance, localization and mapping, and segmentation [46], [55], [1], [25], [26], [16], [39], [89], missing a target in a scene is costly. In these cases, KL and IG metrics are appropriate because they are highly sensitive to false negatives and can assign an arbitrarily large penalty to a saliency map that lacks density at target locations.

AUC can also be used for evaluating saliency for detection applications, being a natural signal detection measure. It is useful for evaluating how confidently (i.e., under what threshold) a saliency map can detect the target fixation locations. The best saliency maps under the AUC metric are ones that assign the highest value to all the target locations while disregarding smaller-valued predictions.

Re-targeting and Compression:

Where it may be important to have different values or level sets denote image regions of varied importance, one of SIM, CC, or KL may be used. Applications including adaptive image and video compression and progressive transmission [28], [32], [50], [85], thumbnailing [52], [72], content-aware image re-targeting and cropping [3], [4], [65], [68], [83], rendering and visualization [38], [48], collage [29], [84] and artistic rendering [18], [37] require ranking (by importance or saliency) different image regions. Under AUC or NSS, a saliency map with a single level set that captures all the target locations

would receive the highest score. Under AUC, saliency maps that place different amounts of density at the correct (fixated) locations will receive a similar score (Fig. 12d). However in re-targeting and compression applications, unlike in detection applications, the relative saliency values of different image regions are important.

Image retrieval and partial matches:

Consider an image retrieval application, where the goal is to find images with a similar distribution of salient or important regions [78], [81], [82]. We may be interested in finding images with similar saliency maps or observers with similar attention patterns (fixation maps) on some set of images. In other words, the goal is to use eye movements or saliency to retrieve images with similar content to a query image.

Histogram intersection is appropriate for handling partial matches, which for image retrieval applications permits the retrieval of candidates that contain the query but may also contain additional content [66], [73]. Partial matches are useful when only part of the image may be known or be of interest, or for handling occlusions and clutter. Analogously in the case of saliency, the SIM metric will highly rank saliency maps that explain the target salient locations at the potential cost of false positives. SIM is a suitable evaluation metric for the task of image retrieval or partial image matching.

Image matching:

In the case of a full image match or retrieval based on saliency, where false positives and false negatives are to be treated symmetrically, a metric like CC may be more suitable. A high similarity score under CC only occurs when the two maps being compared have high and low values in the same locations in an image. CC is also suitable for discovering images that are correlated in the distribution of salient regions, and for discovering subsets of people (or individuals) with correlated eye movement patterns (fixation maps) on the same image.

Robust matching and clustering:

When shifts in feature locations are possible and a more robust image match is desired, EMD is an appropriate metric of choice. In handling false positives, EMD penalizes extra density that is not consistent with nearby regions in the target map.

Additionally, whereas other metrics are capped at the minimum score as soon as there is no overlap between two saliency maps, EMD scores gradually degrade as two maps diverge. This means that EMD can provide a finer-grained comparison between saliency maps, and by capturing the image dissimilarities more closely can be used for clustering and multi-dimensional scaling applications [66].

7 CONCLUSION

We have provided an analysis of the behavior of 8 saliency evaluation metrics. Different evaluation metrics operate on different assumptions: how the ground truth is represented; whether center bias is accounted for by the model; whether false positives or false negatives are more costly; and whether robustness to monotonic transformations and spatial deviations is necessary. Computational considerations also include the type of normalization applied to inputs, and whether the computations are local and differentiable.

Building off the differences in metric behaviors and specific computational properties, we made recommendations for different saliency applications. While other choices are possible, for comparing model performances on standard fixation prediction datasets [14], we recommend Information Gain (IG) and one of Normalized Scanpath Saliency (NSS) or Pearson's Correlation Coefficient (CC). NSS and CC are the metrics most balanced in their treatment of false positives and false negatives, and otherwise provide similar rankings of saliency models as Area under ROC curve (AUC), Similarity (SIM), and Earth Mover's Distance (EMD). Choosing between NSS and CC also requires making a decision about whether the appropriate representation for ground truth is fixation locations or continuous fixation maps, respectively. Continuous fixation maps are more robust than fixation locations (under measurement errors and fewer observers), but require choosing distribution parameters, which may affect reported scores. The Information Gain (IG) metric was recently introduced by Kümmerer et al. [41], [42] to handle center bias. IG measures how much of the ground truth a model can predict above a center baseline, whereas NSS and CC measure model performance at predicting overall viewing behavior including viewing biases. Due to differing modeling assumptions that lead to center bias being accounted for by some models but not others, we recommend complementing NSS or CC with IG.

Under other assumptions or saliency applications, a different selection of metrics may be more appropriate. For instance, metrics that evaluate a model's ability to predict fixation locations (such as AUC and IG) are better suited for detection applications. Under IG and KL, false negatives (missed detections) are especially costly. However, where it is important to evaluate the relative importance of different image regions, placing more or less weight (density) on different locations, distribution-based metrics like SIM and CC might be more favorable. Examples of applications include image-retargeting, compression, and progressive transmission. For image retrieval applications (e.g. retrieving an image according to a distribution of features such as saliency values), SIM and EMD can provide greater robustness, to partial matches and spatial shifts, respectively.

Code for evaluating and visualizing the metric computations is available⁶ to add greater transparency to model evaluation and to allow researchers a finer-grained look into metric computations, to debug saliency models and visualize the aspects of saliency models driving or hurting performance.

REFERENCES

- [1] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk. Salient region detection and segmentation. In A. Gasteratos, M. Vincze, and J. Tsotsos, editors, *Computer Vision Systems*, volume 5008 of *Lecture Notes in Computer Science*, pages 66–75. Springer Berlin / Heidelberg, 2008.
- [2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE Computer Vision and Pattern Recognition*, pages 1597–1604, june 2009.
- [3] R. Achanta and S. Süssstrunk. Saliency detection for content-aware image resizing. In *IEEE International Conference on Image Processing*, pages 1005–1008, 2009.
- [4] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3):10, 2007.
- [5] M. Bindermann. Scene and screen center bias early eye movements in scene viewing. *Vision Research*, 50:2577–2587, 2010.
- [6] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.
- [7] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2012.
- [8] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013.
- [9] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *IEEE International Conference on Computer Vision*, 2013.
- [10] N. Bruce and J. Tsotsos. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162. MIT Press, Cambridge, MA, 2006.
- [11] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 2009.
- [12] N. D. B. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision research*, 2015.
- [13] Z. Bylinskii, E. M. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J. K. Tsotsos. Towards the quantitative evaluation of visual attention models. *Vision research*, 116:258–268, 2015.
- [14] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. MIT Saliency Benchmark. <http://saliency.mit.edu/>.
- [15] R. L. Canosa, J. Pelz, N. R. Mennie, and J. Peak. High-level aspects of oculomotor control during viewing of natural-task images. *Proceedings of SPIE*, pages 240–251, 2003.
- [16] C-K Chang, C. Siagian, and L. Itti. Mobile robot vision navigation & localization using gist and saliency. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4147–4154, 2010.
- [17] A. D. F. Clarke and B. W. Tatler. Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, 102:41–51, 2014.
- [18] D. DeCarlo and A. Santella. Stylization and abstraction of photographs. *ACM Transactions on Graphics (TOG)*, 21(3):769–776, 2002.
- [19] W. Einhäuser and P. König. Does luminance-contrast contribute to a saliency for overt visual attention? *European Journal of Neuroscience*, 17:1089–1097, 2003.

6. <http://saliency.mit.edu/downloads.html>

Metric	Quick take-aways
Area under ROC Curve (AUC)	Historically the most commonly-used metric for saliency evaluation. Driven by high-valued predictions and largely ambivalent of low-valued predictions (including false positives) and monotonic transformations. Good for detection applications.
Shuffled AUC (sAUC)	Introduced to counter the center bias of AUC by scoring a center prior at chance. May have unfavorable behavior especially where the ground truth itself is center biased.
Similarity (SIM)	An easy and fast similarity computation between histograms or distributions. Sensitive to distribution parameters. More sensitive to false negatives than false positives. Applicable to finding partial matches and performing approximate retrieval by query.
Pearson's Correlation Coefficient (CC)	Computes the linear correlation between the prediction and ground truth distributions. Treats false positives and false negatives symmetrically. Can be used for finding correlated saliency maps across images or observers, for matching and retrieval.
Normalized Scanpath Saliency (NSS)	A discrete approximation of CC that is additionally parameter-free (no distribution parameters needed for ground truth), but requires ground truth to be robust. Similar behavior and applications as CC.
Earth Mover's Distance (EMD)	Unlike other metrics, scales with spatial distance. Can thus provide a finer-grained comparison between saliency maps and more robust matching/retrieval. Most computationally intensive, non-local, non-differentiable.
Kullback-Leibler divergence (KL)	Has a natural interpretation where goal is to approximate a target distribution. Highly sensitive to false negatives. Can be used for detection applications.
Information Gain (IG)	Similar information-theoretic formulation as KL, but can subtract what is already explainable by center bias. Can be extended to measure model performance above other baselines. Parameter-free but requires robust ground truth.

TABLE 9: A brief overview of the metric analyses and discussions provided in this paper, highlighting some of the key properties, features, and applications of different evaluation metrics.

- [20] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14), 2008.
- [21] M. Emami and L. L. Hoberock. Selection of a best metric and evaluation of bottom-up visual saliency models. *Image and Vision Computing*, 31(10):796–808, 2013.
- [22] U. Engelke, H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H-J Zepernick, and A. Maeder. Comparative study of fixation density maps. *IEEE Transactions on Image Processing*, 22(3):1121–1133, 2013.
- [23] E. Erdem and A. Erdem. Visual saliency estimation by non-linearly integrating features using region covariances. *Journal of Vision*, 13(4):1–20, 2013.
- [24] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [25] S. Frintrop. General object tracking with a component-based target descriptor. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4531–4536, 2010.
- [26] S. Frintrop, P. Jensfelt, and H. Christensen. Simultaneous robot localization and mapping based on a visual attention system. In *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, pages 417–430. Springer, 2007.
- [27] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *Neural Information Processing Systems*, 2007.
- [28] W. S. Geisler and J. S. Perry. A real-time foveated multiresolution system for low-bandwidth video communication. In *Proc. SPIE: Human Vision and Electronic Imaging*, volume 3299, pages 294–305, 1998.
- [29] S. Goferman, A. Tal, and L. Zelnik-Manor. Puzzle-like collage. *Computer graphics forum*, 29(2):459–468, 2010.
- [30] D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*. John Wiley, 1966.
- [31] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, 2006.
- [32] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, 2004.
- [33] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*, pages 547–554, Cambridge, MA, 2006. MIT Press.
- [34] T. Judd. *Understanding and predicting where people look in images*. PhD thesis, Massachusetts Institute of Technology, 2011.
- [35] T. Judd, F. Durand, and A. Torralba. Fixations on low-resolution images. *Journal of Vision*, 11(4), 2011.
- [36] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.
- [37] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, 2009.
- [38] Y. Kim and A. Varshney. Saliency-guided enhancement for volume visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):925–932, 2006.
- [39] D. Klein, S. Frintrop, et al. Center-surround divergence of feature statistics for salient object detection. In *IEEE International Conference on Computer Vision*, pages 2214–2219, 2011.
- [40] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.
- [41] M. Kümmerer, T. Wallis, and M. Bethge. How close are we to understanding image-based saliency? *arXiv preprint arXiv:1409.7686*, 2014.
- [42] M. Kümmerer, T. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015.
- [43] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):802–817, 2006.
- [44] J. Li, M. D. Levine, X. An, X. Xu, and H. He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on PAMI*, 35(4), 2012.
- [45] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen. A data-driven metric for comprehensive evaluation of saliency models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 190–198, 2015.
- [46] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *IEEE Computer Vision and Pattern Recognition*, pages 1–8, june 2007.
- [47] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.
- [48] P. Longhurst, K. Debattista, and A. Chalmers. A gpu based saliency map for high-fidelity selective rendering. In *Proceedings of the 4th international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*, pages 21–29. ACM, 2006.
- [49] F. Lopez-Garcia, X. Ramon Fdez-Vidal, X. Manuel Pardo, and R. Dosil. Scene recognition through visual attention and image features: A comparison between sift and surf approaches. In Tam Phuong Cao, editor, *Object Recognition*, pages 185–200. InTech, 2011.
- [50] Y-F Ma, X-S Hua, L. Lu, and H-J Zhan. A generic framework

- of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005.
- [51] N. H. Mackworth and A. J. Morandi. The gaze selects informative details within pictures. *Perception & Psychophysics*, 2(11):547–552, 1967.
- [52] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *IEEE 12th International Conference on Computer Vision*, pages 2232–2239, 2009.
- [53] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavioral Research Methods*, 45(1):251–266, 2013.
- [54] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483–2498, 2007.
- [55] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2049–2056, 2006.
- [56] D. Noton and L. Stark. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision research*, 11(9):929–IN8, 1971.
- [57] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107 – 123, 2002.
- [58] D. J. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16:125–154(30), 2003.
- [59] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *European Conference on Computer Vision*, 2008.
- [60] O. Pele and M. Werman. Fast and robust earth mover’s distances. In *IEEE International Conference on Computer Vision*, 2009.
- [61] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005.
- [62] J. Puzicha, T. Hofmann, and H. M. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 267–272, 1997.
- [63] L. W. Renninger, J. Coughlan, P. Verghese, and J. Malik. An information maximization model of eye movements. In *Advances in neural information processing systems*, pages 1121–1128, 2004.
- [64] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *IEEE International Conference on Computer Vision*, 2013.
- [65] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM Transactions on Graphics (SIGGRAPH)*, 2008.
- [66] Y. Rubner and C. Tomasi. *Perceptual metrics for image database navigation*. Springer Science + Business Media, LLC, 2001.
- [67] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:2000, 2000.
- [68] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI ’06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780, New York, NY, USA, 2006. ACM.
- [69] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):1–27, 2009.
- [70] A. K. Sinha and K.K. Shukla. A study of distance metrics in histogram based image retrieval. *International Journal of Computers & Technology*, 4(3):821–830, 2013.
- [71] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
- [72] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 95–104. ACM, 2003.
- [73] M. J. Swain and D. H. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.
- [74] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.
- [75] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643 – 659, 2005.
- [76] A. Toet. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Transactions on PAMI*, 33(11):2131–2146, 2011.
- [77] A. Torralba, A. Oliva, M. S. Castelhan, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766–786, October 2006.
- [78] A. Toshev, J. Shi, and K. Daniilidis. Image matching via saliency region correspondences. In *IEEE Transactions on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
- [79] P-H Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, 9(7):4, 2009.
- [80] N. Vasconcelos. On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Transactions on Information Theory*, 50(7):1482–1496, 2004.
- [81] K. Vu, K. Hua, W. Tavanapong, et al. Image retrieval based on regions of interest. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):1045–1049, 2003.
- [82] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395 – 1407, 2006. Brain and Attention, Brain and Attention.
- [83] D. Wang, G. Li, W. Jia, and X. Luo. Saliency-driven scaling optimization for image retargeting. *The Visual Computer*, 27(9):853–860, 2011.
- [84] J. Wang, L. Quan, J. Sun, X. Tang, and H-Y Shum. Picture collage. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 347–354, 2006.
- [85] Z. Wang, L. Lu, and A. C. Bovik. Foveation scalable video coding with automatic fixation selection. *IEEE Trans. Image Processing*, 12:243–254, 2003.
- [86] N. Wilming, T. Betz, T. C. Kietzmann, and P. König. Measures and limits of models of fixation selection. *PLoS ONE*, 6, 2011.
- [87] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):1–20, 2014.
- [88] A. L. Yarbus. *Eye movements and vision*. Plenum, New York, 1967.
- [89] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. Berg. Studying relationships between human gaze, description, and computer vision. In *IEEE Transactions on Computer Vision and Pattern Recognition*, pages 739–746, 2013.
- [90] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.
- [91] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011.
- [92] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003.

APPENDIX

IMPLEMENTATION AND NOTATIONAL DETAILS

Eye tracking set-up:

Eye movements for the MIT300 dataset were collected using a table-mounted, video-based ETL 400 ISCAN eye tracker which recorded observers' gaze paths at 240Hz. The average calibration error was less than one degree of visual angle. Each image was presented at a maximum dimension of 1024 pixels and the second dimension between 457-1024 pixels (mode: 768 pixels). Images were separated by a 500 ms fixation cross. During pre-processing, the first fixation on each image was thrown out to reduce the center-biasing effects of the fixation cross.

Location-based versus distribution-based metrics:

It is important to note that it is the particular *implementations* of the metrics we use that are either location-based or distribution-based. For instance, there are implementations of AUC and NSS that take the ground truth in as a distribution [53]. The ground truth distribution is first pre-processed into a binary map by being thresholded at a fixed, often arbitrary value. This requires an additional parameter for the metric computation. The (parameter-free) location-based implementations of AUC and NSS are more commonly used for saliency evaluation.

AUC-Judd:

For the AUC-Judd implementation, the ROC curve is constructed by sampling thresholds at all distinct saliency map values. To ensure that enough threshold samples are taken, the saliency map is first jittered by adding a tiny random value to each pixel, thus preventing large uniform regions of one value in the saliency map.

AUC-Borji and sAUC:

The AUC-Borji score is calculated by repeatedly sampling a new set of negatives in 100 separate iterations and averaging these intermediate AUC computations together. On each iteration, as many negatives are chosen as fixations on the current image.

In the shuffled AUC (sAUC) variant, negatives are sampled at random from 10 other randomly-sampled images in the dataset (as many negatives are sampled as fixations on the current image), and the final score is also obtained by averaging over 100 trials.

A note about naming: Riche et al. [64] refer to shuffled AUC as AUC-Borji, but here we make a distinction between Borji's implementation of AUC with randomly-sampled negatives, and sAUC with negatives sampled from other images [9].

Other AUC implementations:

Our AUC implementations are location-based (as in [10], [27], [31], [37], [75], [77]), but other distribution-based implementations of AUC have also been used

in saliency evaluation. If both inputs (ground truth and saliency) are maps, then a number of different implementation choices are possible [53]. The thresholding for computing the ROC curve can be performed on the ground truth map, the saliency map, or both [22]. In the first two cases, one of the maps is thresholded at different values, while the other map is thresholded at a single, fixed value (e.g. to keep 20% of the pixels [77], [53]).

AUC is non-symmetric, and depending on which map is taken as the reference, different scores will be produced. For this reason, an average over two non-symmetric AUC calculations can be computed, by swapping the two maps being compared [22].

A number of additional AUC implementations have been discussed and compared [9], [64], [79].

CC:

A nonlinear correlation coefficient (Spearman's CC), has also be used for saliency evaluation [53], [64], [76]. Unlike Pearson's CC which takes into account the absolute values of the two distribution maps, Spearman's CC only compares the ranks of the values, making it robust under monotonic transformations.

Relationship between CC and NSS:

Recall that NSS is calculated as:

$$NSS(P, Q^B) = \frac{1}{N} \sum_i \bar{P}_i \times Q_i^B$$

where \bar{P} is the normalized saliency map, Q^B is a binary map of fixations, and N is the number of fixated pixels. If the fixations are sampled instead from a fixation distribution Q^D , then the probability that a particular fixation at pixel i is chosen is just the density Q_i^D . By sampling from Q^D , we can construct the binary map Q^B (since $E(Q_i^B) = P(Q_i^B) = Q_i^D$). Over M sets of samples from Q^D :

$$E[NSS(P, Q^B)] = \frac{1}{M} \sum_i \bar{P}_i \times Q_i^D$$

Note that CC can be written as:

$$CC(P, Q^D) = \frac{1}{T} \sum_i \bar{P}_i \times \bar{Q}_i^D$$

Where T is the total number of pixels in the image, and both P and Q^D are normalized. Recall that NSS and CC both normalize by variance. Thus, NSS can be viewed as a kind of discrete approximation to CC.

KL:

The standard implementation of KL that we use is non-symmetric by construction. A symmetric extension of KL can be computed as: $KL(P, Q) + KL(Q, P)$ (also see Jeffrey divergence [62]). Note that this variant would equally penalize false negatives and false positives. However, in this case, the resulting score would not have the natural

interpretation of measuring how good a saliency map prediction is at approximating the ground truth distribution. There is also a shuffled implementation of KL available [90] to discount central predictions.

Relationship between KL and IG:

This relationship is discussed at length by Kümmerer et al. [41], [42]. Here we explicate this relationship for our formulation of KL and IG. Recall that:

$$KL(P, Q^D) = \sum_i Q_i^D \log \left(\epsilon + \frac{Q_i^D}{\epsilon + P_i} \right)$$

Where i iterates over all the pixels in the distribution Q^D (approximating an integral). Then:

$$\begin{aligned} & KL(B, Q^D) - KL(P, Q^D) \\ &= \sum_i Q_i^D \left[\log \left(\epsilon + \frac{Q_i^D}{\epsilon + B_i} \right) - \log \left(\epsilon + \frac{Q_i^D}{\epsilon + P_i} \right) \right] \end{aligned}$$

which for very small ϵ approaches:

$$\sum_i Q_i^D \left[\log \left(\frac{\epsilon + P_i}{\epsilon + B_i} \right) \right]$$

yielding the discrete approximation:

$$\frac{1}{N} \sum_i Q_i^B \left[\log \left(\frac{\epsilon + P_i}{\epsilon + B_i} \right) \right]$$

and within a constant factor (due to change of base of log), this is equal to $IG(P, Q^B)$.

In other words, information gain is like KL but baseline-adjusted (recall also that KL is a dissimilarity metric, while IG is a similarity metric, explaining the change of places between P and B). Kümmerer et al. also point out the difference between *image-based* KL which measures the divergence between the predicted and ground truth distributions (as in this paper) and *fixation-based* KL which measures the divergence of the (histograms of) saliency values at fixated and non-fixated locations. Both versions of KL have been used for saliency evaluation under the metric name KL. Thus, the additional distinction between image-based KL and IG is that the latter uses the original fixation locations as ground-truth and does not require estimating distributional parameters.

IG:

For the visualizations in Figs. 2 and 8, we compute a per-pixel value of: $V_i = \log_2(\epsilon + P_i) - \log_2(\epsilon + B_i)$. This value is then modulated by the human fixation distribution Q^D . In red are all pixels where $Q_i^D V_i < 0$, and in blue are all pixels where $Q_i^D V_i > 0$.

EMD:

We use a fast implementation of EMD provided by Pele and Werman [59] [60]⁷ but without a threshold. For additional efficiency, we resize both maps to 1/32 of their size after they are first resized to the same

dimensions. The maps are then normalized to sum to one. Despite these modifications, EMD is more computationally expensive to compute than any of the other metrics because it requires joint optimization across all the image pixels.

For visualization, we use the formulation of EMD from Sec. 4.2.4. At pixel i of the visualization (as in Fig. 2h and Fig. 9d) we plot $D_{from} = \sum_j f_{ij} d_{ij}$ in green for all i where $D_{from} > 0$, and at pixel j , we plot $D_{to} = \sum_i f_{ij} d_{ij}$ in red for all j where $D_{to} > 0$. Note that the set of pixels where $D_{from} > 0$ is disjoint from the set of pixels where $D_{to} > 0$, so each pixel is either red or green or neither.

Regarding histogram matching:

Prior to September 2014, the MIT Saliency Benchmark histogram matched saliency maps to a target distribution before evaluation [36]. This was intended to reduce differences in saliency map ranges. However, this had significant effects on model performances, inflating or deflating scores, depending on the model. The decision was made to evaluate saliency maps as-is and to leave any preprocessing to the model submitters. This also makes reporting more transparent, as the scores posted on the website directly correspond to the maps submitted.

Regarding empirical limits of metrics:

Other researchers compute the limit of human consistency as the inter-observer (IO) model where all other observers are used to predict the fixations of the remaining observer [7], [53], [61], [86]. The resulting scores are usually averaged over all or a subset of observers.

Center prior model:

This model is created by stretching a symmetric Gaussian to the aspect ratio of the image, so that each pixel's saliency value is a function of its distance from the center (higher saliency closer to center). This version of the center prior performs slightly better than an isotropic Gaussian because objects of interest tend to be spread along the longer axis. See Clarke and Tatler [17] for an analysis of different types of center models.

Single observer model:

We use the fixation map from one observer to predict the fixations of the remaining observers (1 predicting N-1). We repeat this leave-one-out procedure and average the results across all observers. This baseline model indicates how well a single observer predicts an average of observers. Note that it is different from the IO model described above, where all but one observer are used to predict the remaining observer (N-1 predicting 1).

7. Code at <http://www.cs.huji.ac.il/~ofirpele/FastEMD/>